Project title: Multi-Owner data Sharing for Analytics and Integration respecting Confidentiality and OWNer control

Project acronym: MOSAICrOWN

Funding scheme: H2020-ICT-2018-2

Topic: ICT-13-2018-2019

Project duration: January 2019 – December 2021

D2.1

Requirements from the Use Cases

Editors: Rigo Wenning (GEIE ERCIM)
Daniel Bernau (SAP SE)
Reviewers: Stefano Paraboschi (UNIBG)
Sabrina De Capitani di Vimercati (UNIMI)

Abstract

The use cases in MOSAICrOWN correspond to the core problems and business strategies of players in the data market. This deliverable provides the requirements that are representative of the needs of the application domains as scoped by the strategic use cases. It takes into account different techniques to protect the data while paying attention that those complement each other. The requirements follow the data work flow and include the various features that MOSAICrOWN offers to ease targeted and controlled data sharing.

Type	Identifier	Dissemination	Date
Deliverable	D2.1	Public	2019.12.31

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 825333.

MOSAICrOWN Consortium

1.	Università degli Studi di Milano	UNIMI	Italy
2.	EMC Information Systems International	EISI	Ireland
3.	Mastercard Europe	MC	Belgium
4.	SAP SE	SAP SE	Germany
5.	Università degli Studi di Bergamo	UNIBG	Italy
6.	GEIE ERCIM (Host of the W3C)	W3C	France

Disclaimer: The information in this document is provided "as is", and no guarantee or warranty is given that the information is fit for any particular purpose. The below referenced consortium members shall have no liability for damages of any kind including without limitation direct, special, indirect, or consequential damages that may result from the use of these materials subject to any liability which is mandatory due to applicable law. Copyright 2019 by Università degli Studi di Milano, EMC Information Systems International, Mastercard Europe, SAP SE, Università degli Studi di Bergamo, European Research Consortium for Informatics and Mathematics.

Versions

Version	Date	Description
0.1	2019.11.27	Document ready
0.2	2019.12.18	Revised document
1.0	2019.12.31	Final version

List of Contributors

This document contains contributions from different MOSAICrOWN partners. Contributors for the chapters of this deliverable are presented in the following table.

Chapter	Author(s)
Executive Summary	Daniel Bernau (SAP SE), Jonas Boehler (SAP SE)
Chapter 1: Introduction	Daniel Bernau (SAP SE)
Chapter 2: Use Cases	Andrew Byrne (EISI), Michele Mazzola (MC), Tomasz Pawlowicz (MC), Daniel Bernau (SAP SE), Jonas Boehler (SAP SE), Benjamin Weggenmann (SAP SE)
Chapter 3: Requirements Analysis	Daniel Bernau (SAP SE)
Chapter 4: Use Case Requirements Comparison to Current Technologies	Andrew Byrne (EISI), Niamh O'Mahony (EISI), Michele Mazzola (MC), Tomasz Pawlowicz (MC), Daniel Bernau (SAP SE), Jonas Boehler (SAP SE), Benjamin Weggenmann (SAP SE)
Chapter 5: Conclusions	Daniel Bernau (SAP SE), Jonas Boehler (SAP SE)

Contents

List of Figures

List of Tables

Executive Summary

The uptake for data markets is hindered by a lack of trusted, secure platforms for sensitive personal data and confidential company data, for enabling privacy-preserving data sharing and analytics with state-of-the-art privacy and security guarantees. MOSAICrOWN aims to address this issue by developing a comprehensive multi-owner platform providing protection mechanisms in the form of data sanitization, wrapping, and policies.

This deliverable describes "Requirements from the Use Cases" which is the M1-M12 activity for Task T2.1 ("Requirements identification and analysis"). The results of this deliverable are the basis for the development of the technical core of MOSAICrOWN (WP3–5) to allow alignment with research and develop techniques and tools to protect sensitive data. The development will be continuously monitored to align with research activities (T2.2), deployed in a testbed (T2.3) and, finally, validated with regards to the use cases (T2.4) in the course of MOSAICrOWN.

We consider three enforcement phases for each use case: *ingestion*, the process of loading the data into the platform; *storage*, how the ingested data are stored and managed on the platform; *data analytics*, when data of participant are processed in the data market to gain insights from them. To protect the data during these phases different techniques are applied: *data sanitization*, a non-reversible transformation (e.g., differential privacy); *wrapping*, typically reversible if provided a key (e.g., encryption); and *policies* to control data access, usage and sharing.

Deliverable D2.1 presents the actors and activities behind each of the MOSAICrOWN use cases of a multi-owner platform for data markets. Each use case identified by the industry partners describes a real-world scenario in the industry partners' respective field enabled by MOSAICrOWN's data market platform for which no comprehensive solution currently exists. Use Case 1 from EISI focuses on the protection of sensitive data in Intelligent Connected Vehicles (ICV) and considers all protection techniques during the ingestion phase, i.e., transferring sensitive data from a vehicle to the data market in a privacy-preserving fashion. Use Case 2 from MC considers a data sharing ecosystem for financial data where wrapping and policy protections are applied during all enforcement phases. Use Case 3 from SAP SE describes privacy-preserving consumer analytics (for operational data and experience data, e.g., historical purchases) enabled by data sanitisation techniques in all three enforcement phases.

We illustrate that the use cases complement each other along the data enforcement dimensions – ingestion, storage, and data analytics – by analyzing sanitization, wrapping and policy requirements per use case. A detailed analysis of functional and non-functional requirements, and respective dependencies, is presented per use case. Given the individual requirements, a synthesis of the individual use case requirements is presented to identify potential synergies or mutual interests. Furthermore, we present the specific connection of each of the use cases to the state-of-the-art.

1. Introduction

Academia and industry meet within MOSAICrOWN to demonstrate through actual business use cases, effective data protection techniques in multi-owner platforms for data markets. The result of MOSAICrOWN will be then a set of solutions providing for an enriched data market scenario (Figure 1.1). The novel solutions that will be developed in the project shall enhance the trust of end-users (i.e., data owners), and enable the increased adoption of data sanitization and wrapping concepts for privacy and confidentiality when sharing personal or sensitive data. The objectives of the MOSAICrOWN project are aligned with the stated goals of the European Commission: providing effective data protection and governance in multi-owner platforms for end-users, demonstrating a rich set of protection techniques within operational environments of industrial partners, and strengthening the competitive position of European industry and the European leadership in promoting and sustaining data protection.

To meet these objectives MOSAICrOWN provides usable techniques enabling data owners (from large companies to individual users) to control and specify policies regulating protection of their data and apply data sanitization or wrapping for storage and/or data analytics. Such protection guarantees are also offered to data owners as well as data consumers accessing such data and services. MOSAICrOWN will therefore significantly decrease the lack of trusted and secure platforms, and privacy-aware analytics methods for secure sharing of personal or sensitive data. To transfer the advanced techniques towards concrete industrial applications, MOSAICrOWN techniques are provided as a modular set of services, with practical applicability and compatibility with current techniques and interfaces.

1.1 Purpose of this Document

The initial goal of Work Package 2 in months 1 – 12 is to produce, identify and analyze the data protection requirements of the different use cases, and detail the needs for data governance, sanitization, and wrapping. Each use case addresses a specific application scenario, which can be classified as IoT data sharing, data markets for financial data, and sanitization capabilities for experience and organizational data. This document (D2.1) is the first output that informs the technical core WPs in terms of the MOSAICrOWN use-case requirements. For each use case, this document identifies the actors, their actions and describes the detailed requirements of the services central to these use cases. Requirements are also determined by the legal and business environment surrounding likely deployments. The requirements will be iteratively refined, steering the entire technological development process according to the project goals and objectives.

1.2 Structure of the Document

The rest of this deliverable is structured as follows. Chapter 2 provides an overview of each of the use cases. Chapter 3 considers the specific use case requirements, and seeks to categorize them

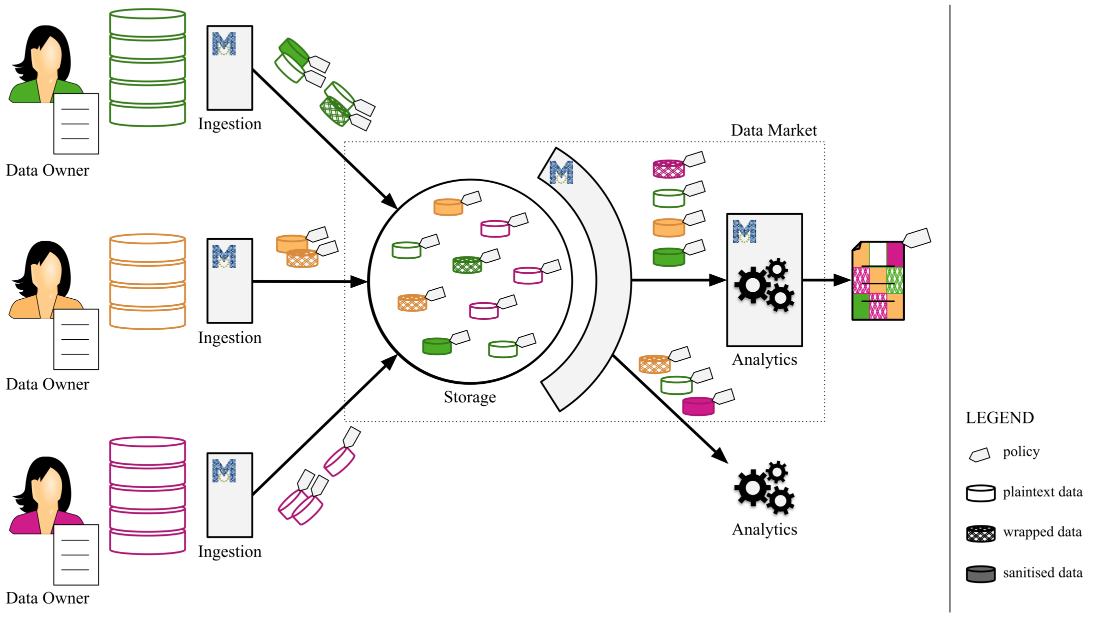

Figure 1.1: MOSAICrOWN's enriched data market

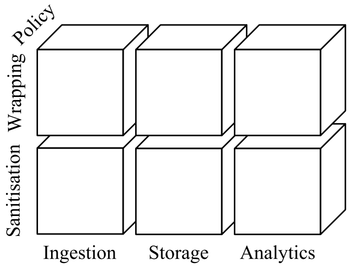

Figure 1.2: MOSAICrOWN dimensions

into a common set of requirements for MOSAICrOWN. Chapter 4 considers current technologies used in data markets for providing data confidentiality and sanitization. It takes these as a starting point, and examines how MOSAICrOWN could enhance them. Chapter 5 states the findings of this deliverable and draws conclusions for the subsequent work within MOSAICrOWN.

1.3 MOSAICrOWN Dimensions

MOSAICrOWN provides data protection techniques that span the complete data life-cycle within data markets. Each of the MOSAICrOWN use cases evaluates given techniques from a number of perspectives, which we refer to as MOSAICrOWN *dimensions*. The MOSAICrOWN dimensions are *a)* requirements capturing (i.e., policies); *b)* enforcing technologies (i.e., wrapping and sanitization); *c)* enforcement phase (i.e., ingestion, storage, analytics). Figure 1.2 illustrates the dimensions of MOSAICrOWN. Not all dimensions are applicable to all use cases, and likewise, not all use cases are applicable to all dimensions to the same extent. However, the broad range of protection aspects covered by the use cases ensures that all the challenges highlighted by the dimensions are explored.

1.4 Requirements Analysis Methodology

The methodology for use case requirements analysis starts with a description of the underlying business scenarios. These scenarios cover the key elements of the problem being addressed, and thus the core features of the project which will be implemented/validated in the use cases. For each use case, our methodology describes the entities that are expected to interact with the system. They can be people or other parts of the system. Some actors may be common to more than one use case. In some cases an actor maps well to a stakeholder in the project, and this is noted. Actors cause events at data market (e.g., ingest data, define policies or analyze data). Thus, we incorporate an analysis of sequences per use case in the form of a written description accompanied by a diagram showing the interactions between the actors and the system. The sequence is intended to flow similar to describing a story (e.g., for a demo).

Within this deliverable, we differentiate requirements into two categories: functional requirements and non-functional requirements. Functional requirements define the expected behavior and functions of a system and are further detailed after initial identification within the system design phase. System design will happen throughout the deliverables in WP2. Example areas for technical requirements are security, accuracy or interoperability. Non-functional requirements address the operational needs of a system. These requirements, for example, address maintainability, usability or portability.

Requirements within this document are concise, complete, unambiguous, verifiable, and necessary. Each requirement is uniquely identified with a meaningful tag allowing an overall requirements catalogue to be derived. A tag has the structure REQ-UC*x*-*yn*, where $x \in \{1,2,3\}$ identifies the Use Case (UC), y groups the requirements by dimension, phase or topic (e.g., DI for data ingestion) and n is a number (as more than one requirement per group y is possible). This deliverable addresses three objectives: *1)* identify functional and non-functional use case requirements; *2)* synthesize use case requirements into a common set of requirements; *3)* consider the challenges posed by the use cases, and the available data governance, sanitization and wrapping technologies, and explain how MOSAICrOWN will help address these challenges for the data market. Common requirements provide a shared set of terms and concepts across MOSAICrOWN, and provide input from real scenarios that guide the research of the core technical work packages. The methodology used to synthesize the common requirements consists in *1)* classifying specific use case requirements by the MOSAICrOWN dimensions; *2)* subdividing the specific requirements into logical categories, based on a synthesized common set of concepts; *3)* elaborating common MOSAICrOWN requirements, specifically stated or implied, from the categorizations and the specific use case requirements.

The comparison to current technologies is performed by each use case owner by reviewing their specific use case requirements, and comparing them to current data governance, sanitization and wrapping industry techniques and research. The analysis yields the current shortcomings and explains how MOSAICrOWN solutions will help address them.

2. Use Cases

This chapter illustrates the relevance of the MOSAICrOWN dimensions to each use case scenario and the requirements for the different use cases.

2.1 Use Case 1: Protection of Sensitive Data in an Intelligent Connected Vehicle (ICV) Data Market (EISI)

Use Case 1 involves multiple parties bringing together disparate data sources in the context of intelligent connected vehicles (ICV) for the provision of monetized services that will create new data markets in an emerging sector. The automotive industry is undergoing a period of massive transformation as smart, connected technologies transfer to the domain of connected cars and autonomous driving. In parallel to this, many sectors are undergoing similar transformations that will reshape how services are provided, how infrastructure is planned, and how resources are managed. For example, data collected from connected vehicles are valuable to municipalities as they can derive insights into traffic congestion that will aid future planning, or detect potholes in real-time to alert road maintenance crews.[1] Data collected from connected vehicles will create new data markets in insurance, where insurance providers will access data supplied by sensors on the car to generate tailored insurance policies.

The common component behind these new opportunities during these transformations is *data*. Data are the enabler of new markets that will uncover the customer, service, and operational insights, which will ultimately create new monetization opportunities. However, with increasing data regulations backing up the need to use data in an ethical way, innovation risks being stifled. The scenario proposed in Use Case 1 will demonstrate the effectiveness of MOSAICrOWN data protection mechanisms in real situations where a viable ICV data market requires the application of a data governance model and supporting data wrapping and sanitization, to provide secure, granular access to data sharing and analytics. The primary use case that MOSAICrOWN will implement is illustrated in Figure 2.1. The actors involved in the use case are, the connected vehicle fleet, the electric vehicle (EV) charging infrastructure provider, and the MOSAICrOWN cloud provider hosting the data sharing and analytics platform. Open sources of data available through open data portals[2] will also supplement data collected from connected vehicles to provide additional data for the analytics processes.

In this use case, the connected vehicle fleet manager and the EV charging infrastructure provider want to exchange data such that they can derive mutually beneficial insights into the status of the EV charging infrastructure. For example, from the perspective of the fleet manager, knowledge of EV charge point status (e.g., availability, health, queue information, capacity) can

[1] https://www.jaguarlandrover.com/news/2019/04/money-earn-you-drive-jaguar-land-rover

[2] https://data.gov.ie/data set/ev-charge-points

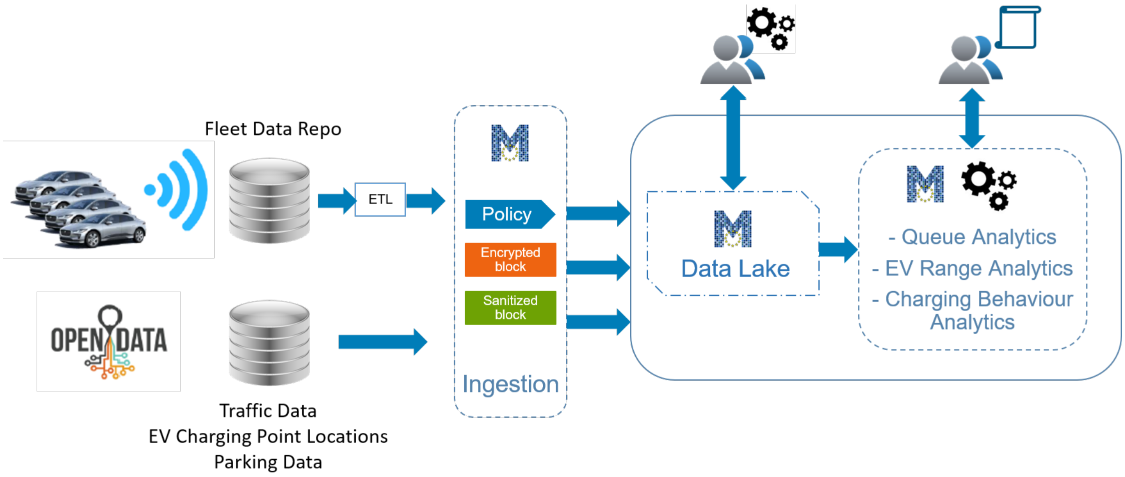

Figure 2.1: ICV use case for fleet charging optimization

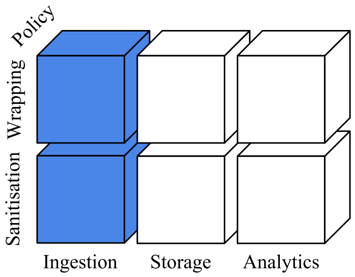

Figure 2.2: Overview of the Use Case 1 dimensions

help inform intelligent decision making that directs vehicles towards nearby charge points. Combined with data collected from the car (e.g., location, battery status, route), drivers can be alerted to divert to recommended charge points that are available. On the other side, the EV charging infrastructure provider can leverage data from their own EV charging infrastructure and from connected electric vehicles using the infrastructure to identify patterns in usage, inform infrastructure planning, and detect faults in the infrastructure. As noted above, restrictions due to data privacy concerns are inhibitors to innovation and the development of new data markets. The data used in this use case prompt these exact challenges, as the data include sensitive information that could lead to the identification of a person such as GPS data, or vehicle identification numbers (VIN). The data are also commercially sensitive as they reveal performance and usage data for EV charging stations, as well as performance related data from the connected vehicles. In this use case, it is also noted that it is not just required to protect the raw data that are collected. The insights gained from the data through analytics are more valuable, and should therefore also have data protection mechanisms for wrapping and sanitization applied to protect against unauthorized access.

Figure 2.2 describes the MOSAICrOWN dimensions that Use Case 1 aims to directly evaluate. The emphasis here is on data ingestion to the platform, taking into account the application of data governance policy, data wrapping techniques, and data sanitization techniques. It is to note, however, that some of the requirements naturally touch, as it is to be expected, also the subsequent phases (i.e., storage and analytics) of the data life-cycle. Determining how close to the data source the data protection perimeter must be for the ingestion phase is an important aspect of the ICV use case. With limited storage on board vehicles, data must be quickly offloaded to edge or cloud services to be processed and stored. Therefore MOSAICrOWN must consider portable solutions that are flexible in their deployment without risk to security.

2.1.1 Functional requirements

Data Ingestion

- REQ-UC1-DI1: MOSAICrOWN ingestion mechanism should support *close to source* deployment.
- REQ-UC1-DI2: Ingestion mechanism should support real-time stream data handling.
- REQ-UC1-DI3: Ingestion mechanism should support batch data handling.
- REQ-UC1-DI4: Ingestion mechanism should support different data types and formats, and account for structured and unstructured data.
- REQ-UC1-DI5: Ingestion mechanism should implement data wrapping and sanitization functions according to the defined data governance model.
- REQ-UC1-DI6: Efficient data compression and encryption functions should be used to reduce payload size.
- REQ-UC1-DI7: Identifiers (e.g., VIN) should be preserved but secured from unauthorized access.
- REQ-UC1-DI8: Sanitization should allow for assessment and notification of data completeness (e.g., identify missing fields and send alert).
- REQ-UC1-DI9: Data sanitization should support protection of sensitive information from linkage attacks using multiple fields from the data sources.
- REQ-UC1-DI10: Ingestion mechanism should support ingestion from multiple concurrent sources.

Data Governance

- REQ-UC1-DG1: A clear language and set of definitions for describing the data governance model are required to support the definition and enforcement of access control rules and data protection mechanisms on the data.
- REQ-UC1-DG2: The platform should provide the capability for data providers to define data governance models for new data sets entering the platform.
- REQ-UC1-DG3: Data protection parameters for wrapping and sanitization should be configurable by the data owner.

Access Control Management

- REQ-UC1-AC1: Access control and authorizations mechanisms should support varying levels of granularity.
- REQ-UC1-AC2: It should be possible to grant and revoke access to specific data sets or fields.
- REQ-UC1-AC3: A centralized key management infrastructure should be implemented to support management of user access. This should be deployed on a separate server to the use case platform.
- REQ-UC1-AC4: The platform should allow restricting some data sets and/or platform users (providers or consumers) to use the platform for sharing only, analytics only, or both analytics and sharing.
- REQ-UC1-AC5: The platform should allow data providers to share data with a particular data consumer.
- REQ-UC1-AC6: The platform should allow data providers to share data with multiple data consumers.
- REQ-UC1-AC7: Policies for data sets and platform users should be configurable by the data provider.

Data Management

- REQ-UC1-DM1: Mechanisms for tracking data movement and access on the platform should be implemented.
- REQ-UC1-DM2: Data should be accessible to all authorized consumers of the data at the same time, regardless of preferred format.
- REQ-UC1-DM3: Integrity of original data should be maintained separately in isolation, allowing an authorized entity to access data in their original state.
- REQ-UC1-DM4: Platform should provide guarantees over data removal.
- REQ-UC1-DM5: Data are protected at rest and in transfer.

Data Processing

- REQ-UC1-DP1: Output from the analytics process can be stored on the platform, according to the data governance specification.
- REQ-UC1-DP2: Output from the analytics process should be sanitized according to the data governance specification.
- REQ-UC1-DP3: Output from the analytics process can feedback into the shared data set, enriching the data.

2.1.2 Non-functional requirements

Data Economy

- REQ-UC1-DE1: The platform should support licensed model.
- REQ-UC1-DE2: Distinction should be clear for consumers of the platform between those requesting *data analytics* functions and *data sharing* functions.

Performance

- REQ-UC1-P1: Operations requiring critical decision making should only incur a relatively small additional latency due to data wrapping and sanitization mechanisms.
- REQ-UC1-P2: Sanitized data should retain sufficient utility to render them valuable for sharing or analysis.

Code Quality

- REQ-UC1-CQ1: Code developed for data protection libraries, and other platform components must follow recommended coding practices.
- REQ-UC1-CQ2: Code coverage for testing should be $> 90\%$.

2.1.3 Requirements catalogue

Req. Ref.	Description	Dimension	Priority	Dependencies on other requirements
REQ-UC1-DI1	Close to source deployment	Ingestion	High	
REQ-UC1-DI2	Real-time stream handling	Ingestion	Medium	
REQ-UC1-DI3	Batch handling	Ingestion	Medium	
REQ-UC1-DI4	Support for different data types, structured and unstructured data	Ingestion	High	
REQ-UC1-DI5	Data wrapping and data sanitization	Ingestion, Policies	High	
REQ-UC1-DI6	Compression and Encryption	Ingestion	Medium	
REQ-UC1-DI7	Secure identifier preservation	Ingestion, Storage, Policies	High	
REQ-UC1-DI8	Assessment of data completeness	Ingestion	Low	
REQ-UC1-DI9	Protection from linkage attacks	Ingestion, Sanitization	High	REQ-UC1-DI5

REQ-UC1-DI10	Support ingestion from multiple concurrent sources	Ingestion	Medium	
REQ-UC1-DG1	Language and definitions for data governance	Ingestion, Storage, Analytics, Policies	Medium	
REQ-UC1-DG2	Support data governance models per data set per data provider	Ingestion, Storage, Analytics, Policies	Medium	REQ-UC1-DG1
REQ-UC1-DG3	Wrapping and sanitization parameters configurable by data owner	Ingestion, Policies	Medium	REQ-UC1-DG1, REQ-UC1-DG2
REQ-UC1-AC1	Access control and authorization	Storage	Medium	
REQ-UC1-AC2	Grant and revoke access	Storage	High	
REQ-UC1-AC3	Centralized key management infrastructure	Ingestion, Storage, Analytics, Policies	Medium	
REQ-UC1-AC4	Support limitation to data sharing or analytics	Storage, Analytics, Policies	High	REQ-UC1-DI5, REQ-UC1-DG2
REQ-UC1-AC5	Allow data sharing between providers and consumers	Storage, Policies	High	REQ-UC1-DI5, REQ-UC1-DG2, REQ-UC1-AC2, REQ-UC1-AC3
REQ-UC1-AC6	Allow data sharing between multiple parties	Storage, Policies	High	REQ-UC1-DI5, REQ-UC1-DG2, REQ-UC1-AC1, REQ-UC1-AC2, REQ-UC1-AC3
REQ-UC1-AC7	Policies configurable by data provider	Storage, Policies	Medium	REQ-UC1-DG1
REQ-UC1-DM1	Tracking data movement and access	Ingestion, Storage, Analytics	High	
REQ-UC1-DM2	Accessibility of data	Storage	High	
REQ-UC1-DM3	Integrity of original data	Ingestion, Storage	Low	

REQ-UC1-DM4	Support for deletion guarantees	Storage	High	
REQ-UC1-DM5	Protection at rest and in transfer	Ingestion, Storage, Analytics	High	REQ-UC1-DI5
REQ-UC1-DP1	Support storing data analytics results	Analytics, Policies	Medium	
REQ-UC1-DP2	Anonymization of data analytics results	Analytics, Sanitization, Policies	Medium	REQ-UC1-DG2, REQ-UC1-DG3
REQ-UC1-DP3	Merge data analytics results with shared data	Storage, Analytics, Policies	Low	REQ-UC1-DP2
REQ-UC1-DE1	License model	Ingestion, Storage, Analytics	Low	
REQ-UC1-DE2	Distinction between data sharing and analytics	Storage, Analytics, Policies	Medium	
REQ-UC1-P1	Limiting latency caused by wrapping and sanitization	Ingestion, Analytics, Sanitization, Wrapping	Medium	REQ-UC1-DI5
REQ-UC1-P2	Ensure utility under sanitization	Analytics, Sanitization, Policies	High	REQ-UC1-DI5
REQ-UC1-CQ1	Consider recommended coding practices	Ingestion, Storage, Analytics, Policies	High	
REQ-UC1-CQ2	Code covering for testing	Ingestion, Storage, Analytics, Policies	Medium	

Table 2.1: Use Case 1 Requirements

2.2 Use Case 2: Data markets for analysis of financial data (MC)

Use Case 2 involves sensitive financial data, customer behavior data, and customer demographic information between multiple Commercial Organizations in the B2B (Business to Business) space (e.g., Financial institutions like banks, Credit unions, Lenders, Payments network operators). The data include information about Financial performance of the business, Financial Products (e.g., Saving accounts), Financial Instruments (e.g., Checks, Debit card, Credit card etc), Payments information (e.g., monthly card balance payments), Purchase behavior data (e.g., Retail and commerce spend, spend at point of sale etc). The safety and security of this information are of paramount importance due to the nature of data, it is both Personally Identifiable Information (PII) as well as financially material (commercially sensitive) information that is governed by

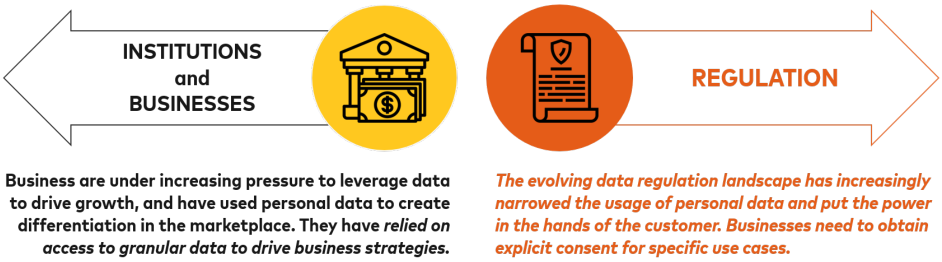
INSTITUTIONS and BUSINESSES
REGULATION
Business are under increasing pressure to leverage data to drive growth, and have used personal data to create differentiation in the marketplace. They have relied on access to granular data to drive business strategies.
The evolving data regulation landscape has increasingly narrowed the usage of personal data and put the power in the hands of the customer. Businesses need to obtain explicit consent for specific use cases.

Figure 2.3: Opposing forces in the market place

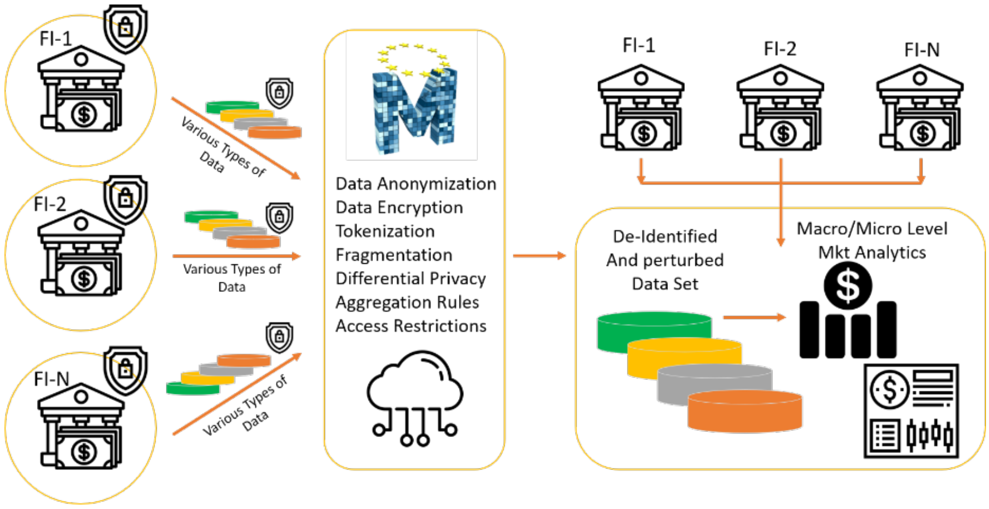

Figure 2.4: Data Sharing Ecosystem

non-Compete clauses and Anti-Trust regulations. The changing regulatory landscape has however created an environment where the business interests and the regulatory restrictions have become the opposing forces (Figure 2.3). In an increasingly competitive market place, the businesses are under pressure to be more data centric and personalize products/offers to the customers needs to stay relevant and engaged with the customers. The evolving data regulation landscape on the other hand has increasingly narrowed the usage of personal data and put the power in the hands of the customer. Businesses now need to obtain explicit consent for specific use cases. Governments, commercial enterprises and charitable organizations have therefore been reluctant to share data within the new regulatory landscape in the advent of GDPR. Sharing of data is critical to the growth of the European economy and its limitation can damage innovation and collaboration.

MC has a business need to operate on combined data provided by the constituent parties to analyze financial transactions to produce market level analytics and forecasts at both microeconomic and macroeconomic levels. With the advent of open banking there is a core need to pool data from all the players in the ecosystem to enable new and improved product offerings as well as analysis on customer needs, and utilization of current products. Figure 2.4 highlights the high-level flow of this data pooling ecosystem.

The platform envisioned and managed by MC offers two modes for access to the analytics function. The first mode assumes that the analytics is executed outside the data market. Controls on the correct protection of data must be executed both in the ingestion phase and in the transfer of data toward the external analytics (represented by the shields in Figure 2.4). The second mode instead assumes that the analytics is executed within the data market, with the application of

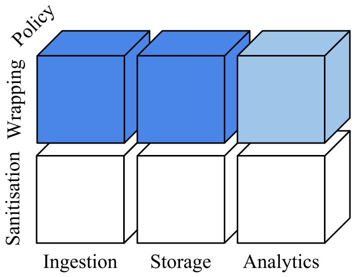

Figure 2.5: Overview of the Use Case 2 dimensions

protection in the execution of the computation and on the communication of the results.

It is important to note, that it is paramount to protect the privacy of individual data contributor's information, but it is also critical that the combined data asset be protected with the same rigor. Since the access to combined information can be used to gain undue market level advantage. Also, there is the need to track the origins and lineage of the data in the data market life-cycle. The development of novel techniques for providing effective data protection will enable better and enriched data analytics.

Figure 2.5 describes the MOSAICrOWN dimensions that apply to Use Case 2. MOSAICrOWN will provide solutions for enabling the sharing and processing of microdata in respect of privacy regulations and on possible privacy/confidentiality constraints holding over the data. It will enable the expression of protection requirements and their representation as policies regulating data access, usage, and sharing.

Use Case 2 requires support of the governance framework for the specification of policies holding over data and their consideration across the data life-cycle. The Use Case will require techniques, provided by data wrapping (developed within WP4), which are first operated by the owner (ingestion phase), and by data processor (i.e., MC itself) at the storage and analytics phase (in case of internal processing) or at the storage phase only (in case of data extraction for external analytics).

The data wrapping layer will enable the data contributor or MOSAICrOWN to determine the relative sensitivity of the data elements provided. Based upon the different levels of sensitivity the system should provide the flexibility to select various different data wrapping techniques to be applied to the data. The variety offered from the system, through a catalogue, will provide a secondary level of security from an outward attack. The system should be built to handle any combination of wrapping techniques such as pseudo-anonymization, encryption, hashing, tokenization, and data fragmentation.

These techniques are important because they ensure the protection of all data assets by processing sensitive data in such a way that the data can no longer be attributed to a specific data subject without the use of additional information. Additionally, the techniques help protect the data in transit and at rest using encryption protections. All of these techniques, in use together, significantly reduce the risks associated with data processing, while maintaining the datas utility.

2.2.1 Functional Requirements

Data Ingestion

- REQ-UC2-DI1: Ingestion mechanism should support batch data handling, which will enable the best level of automation.

- REQ-UC2-DI2: Ingestion mechanism should support different data types and formats (e.g., alphanumeric, integer, floating).

- REQ-UC2-DI3: All data feeds will be provided via governed and managed APIs.

- REQ-UC2-DI4: The platform should evaluate the need to perturb data utilizing K-anonymity (are the segments of individual records big enough). Within this anonymity evaluation, it is important to have a level of diversity (so segments are not too unique (L-diversity)) to protect individual attribute uniqueness.

Access Control Management

- REQ-UC2-AC1: Data set, Table row, Table column level access controls will be implemented with the goal of giving customized privileges to data. This would be to control who could see individual data sets, attributes, and measures.

- REQ-UC2-AC2: User level access control management system will be built into the system. This will allow giving permissions to individual users based upon their roles and what they should or should not see. It also gives the flexibility to the system to add and remove users quickly and efficiently.

- REQ-UC2-AC3: Business/Organization level access control management system will be built into the system. This will allow giving base permissions to an organization as a whole. It also gives the flexibility to the system to delete a large set of users if the organization ceases to have a relationship with the data marketplace.

- REQ-UC2-AC4: All data accesses will be provided via governed and managed APIs.

- REQ-UC2-AC5: Data consumers must have permissible purpose to use the data.

- REQ-UC2-AC6: Data exports will be limited to aggregated data and/or model code. Data exported to be evaluated by statistical test to ensure they are anonymized before any data is available for viewing by any user.

Data Management

- REQ-UC1-DM1: Data controller right to the data need to be protected for MOSAICrOWN. The entity that runs the marketplace becomes the controller of that data. It has the same rights as the original contributor of the data.

- REQ-UC1-DM2: A controller to controller contract will be signed, between the contributor and the entity receiving the data transfer. This step is important because each time noise/perturbation needs to be applied to the data, approval will not be necessary because the data controller will have the required rights to make the decision.

- REQ-UC2-DM3: Repository offers the functionality to combine multiple data sets (ingested from different data providers) that reside in the data marketplace due to the rights obtained by the data controller.

- REQ-UC2-DM4: Checks and balances to limit data combinations to avoid re-identification.

Data Wrapping

- REQ-UC2-W1: The system will be designed to evaluate the data provided by each contributor to determine the relative sensitivity of the data elements provided. These data elements will then be categorized into appropriate buckets to apply appropriate data wrapping techniques.

- REQ-UC2-W2: Based upon the different levels of sensitivity and categorization established in REQ-UC2-W1, the system should enable the selection of various different wrapping techniques that need to be applied to that data (techniques will be determined based upon the level of risk of the data). The goal is to offer flexibility in terms of security and data wrapping approach. This variety in options will provide a secondary level of security from an outward attack.

- REQ-UC2-W3: All of these data wrapping techniques will be outlined in a data catalogue (similar to an a la carte menu) that will be given to the data contributor. The data contributors will be given the option to select which techniques they would like applied to their data down to a column level.

- REQ-UC2-W4: The system should be built to handle any combination of wrapping techniques (pseudo-anonymization, encryption, hashing, tokenization, and data fragmentation).

- REQ-UC2-W5: Mask PII data using pseudonymization so that the data can no longer be attributed to a specific data subject without the use of additional information. The additional information must be kept separately and subject to technical and organization measures to ensure non-attribution to an identified or identifiable person.

- REQ-UC2-W6: Unique Identifiers (e.g., Customer Ids, Account numbers) should be anonymized at the source as well as re-anonymized in the repository.

- REQ-UC2-W7: Encryption algorithms will be applied when the data contributor is submitting data to the data marketplace. This allows the data to be secure in transit as they pass through the internet or from network to network.

- REQ-UC2-W8: The data pipelines should enable the use of multiple encryption algorithms, such as TripleDES, AES, RSA, and Twofish based upon the requirements of the data contributor and data controller.

- REQ-UC2-W9: Encryption keys should be physically stored in different locations than the encrypted data and appropriate security protocols applied to this storage.

- REQ-UC2-W10: The data pipelines should enable the option to apply hashing to sensitive/identifiable data (e.g., PAN, SS#) using one of available hashing algorithms (SHA1, SHA256, MD5) so that the data are not identifiable and cannot be connected back to any one individual.

- REQ-UC2-W11: The data pipelines should enable the option to apply a salt to the hashing algorithm chosen. Adding this additional salt to the input of a hash function will guarantee a unique output and varying this salt by each contributor will help protect the data.

- REQ-UC2-W12: There will be no creation of Rainbow tables that link the hashed and clear values. The clear values of the data will be destroyed at the time of creation of the hashed data.

- REQ-UC2-W13: The system should enable the support of tokenization where sensitive data (e.g., PANs) will be replaced with tokens, a series of randomly generated numbers. The tokenization option will minimize the risk of exposing sensitive data and protect against data breaches.

- REQ-UC2-W14: The data pipeline will store the generated tokens and their corresponding data in an encrypted token vault only accessible by the data controller.

- REQ-UC2-W15: Wherever applicable and desired by the contributor, the system should support horizontal data fragmentation (e.g., sharding). The data set will be split horizontally into rows and stored across multiple database servers, thus enabling data protection because all the information is not stored in the same place and there is no easy way to determine where the other data reside.

- REQ-UC2-W16: Wherever applicable and desired by the contributor, the system should support vertical data fragmentation when storing the data. The data set will be split by the columns and stored across multiple database servers, greatly reducing the risk of leaking the entire database and providing easier query access to the data.

- REQ-UC2-W17: Wherever applicable and desired by the contributor, the system should support hybrid data fragmentation. This gives the flexibility to the data controller to employ a combination of horizontal and vertical fragmentation.

Sanitization

- REQ-UC2-S1: Repository needs to implement the following:
  - Data evaluation during on-boarding to avoid ingesting personal or sensitive data
  - Detection techniques to isolate records likely to contain personal data.
  - Handling sensitive data

- REQ-UC2-S2: Remove unique outliers in data set as data are loaded.

2.2.2 Non-Functional Requirements

Data Controller

- REQ-UC2-DC1: Data Repository needs to be an independent organization and a separate data controller.

- REQ-UC2-DC2: External unique identifiers to be substituted with repository specific unique identifiers for use internally only within repository.

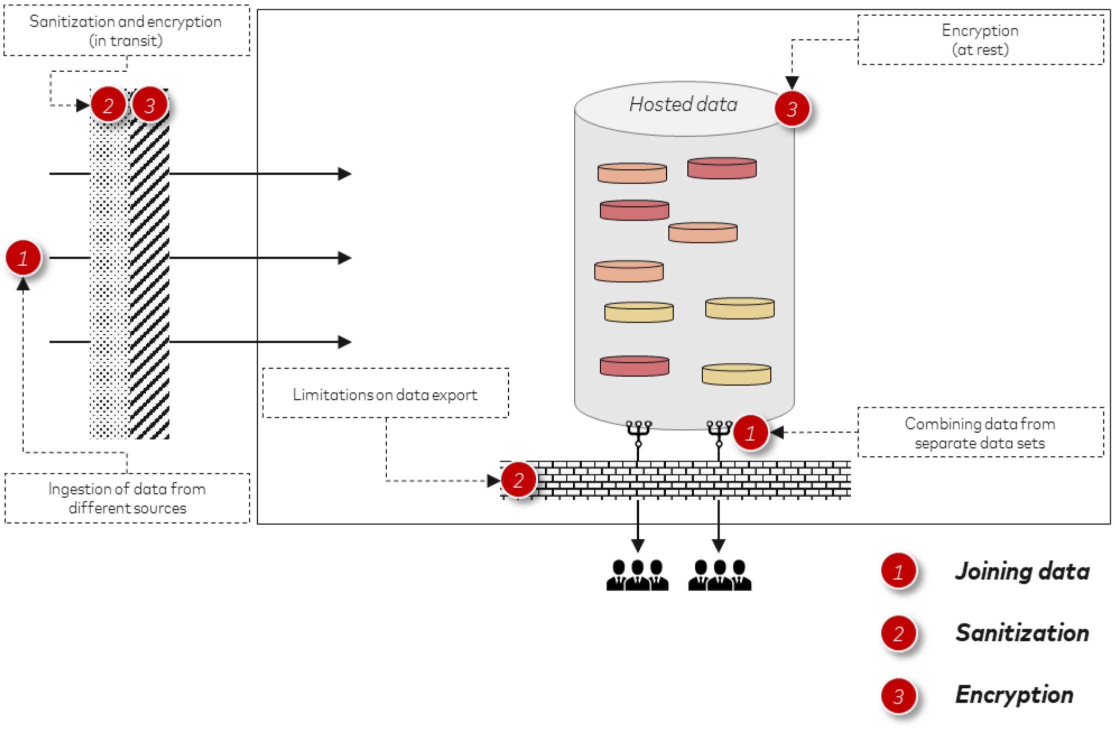

Figure 2.6: Illustration of key functional requirements

Data Analytics

- REQ-UC2-DA1: Standard data science libraries will be provided that help authorized users to analyze combined data for data science purposes.
- REQ-UC2-DA2: The platform will provide a certain level of Service Level Agreements (SLAs) and be designed for analytics and processing on top of large scale data sets with multiple terabytes of data.

Figure 2.6 illustrates functional and non-functional requirements for a data repository in MOSAICrOWN.

2.2.3 Requirements Catalogue

Req. Ref.	**Description**	**Dimension**	**Priority**	**Dependencies on other requirement**
REQ-UC2-DI1	Batch handling	Ingestion	High	
REQ-UC2-DI2	Data types and formats	Ingestion	High	
REQ-UC2-DI3	Data feeds via APIs	Ingestion	Low	
REQ-UC2-DI4	Data perturbation	Ingestion	Medium	
REQ-UC2-AC1	Access control levels	Storage	High	REQ-UC2-DM3

REQ-UC2-AC2	User level access control	Storage	High	
REQ-UC2-AC3	Business/Org level access control	Storage	Low	REQ-UC1-DM1
REQ-UC2-AC4	API Access only	Storage	Medium	
REQ-UC2-AC5	Permissible Purpose	Storage	High	
REQ-UC2-AC6	Data export - aggregates only	Storage	Medium	
REQ-UC1-DM1	Data controller rights	Storage	High	
REQ-UC1-DM2	Controller to Controller Contract	Storage	High	
REQ-UC2-DM3	Data combination and merging	Storage; Analytics	High	
REQ-UC2-DM4	Limit data combinations	Storage; Analytics	Medium	
REQ-UC2-W1	Data Risk Assessment	Data Wrapping	High	REQ-UC2-DM1
REQ-UC2-W2	Data Wrapping Approach	Data Wrapping	High	REQ-UC2-W1
REQ-UC2-W3	Data Catalogue	Data Wrapping	High	
REQ-UC2-W4	Wrapping Techniques	Data Wrapping	High	REQ-UC2-W1
REQ-UC2-W5	Pseudonymization	Data Wrapping	High	REQ-UC2-DI4
REQ-UC2-W6	Re-anonymization for UIDs	Data Wrapping	Low	REQ-UC2-DC2
REQ-UC2-W7	Data linkage protection	Data Wrapping	Low	
REQ-UC2-W8	Encryption Algorithms	Data Wrapping	High	REQ-UC2-DI4
REQ-UC2-W9	Encryption Keys	Data Wrapping	Medium	
REQ-UC2-W10	Hashing	Data Wrapping	High	REQ-UC2-W9
REQ-UC2-W11	Salting	Data Wrapping	High	REQ-UC2-W3
REQ-UC2-W12	Rainbow Tables	Data Wrapping	High	REQ-UC2-W11
REQ-UC2-W13	Tokenization	Data Wrapping	Medium	
REQ-UC2-W14	Token Vault	Data Wrapping	High	
REQ-UC2-W15	Horizontal Data Fragmentation	Data Wrapping	High	REQ-UC2-W14
REQ-UC2-W16	Vertical Data Fragmentation	Data Wrapping	High	
REQ-UC2-W17	Hybrid Data Fragmentation	Data Wrapping	High	
REQ-UC2-S1	Data checks	Sanitization	High	

REQ-UC2-S2	Unique Outlier	Sanitization	Medium	
REQ-UC2-DC1	Data controller	Storage; Analytics	High	
REQ-UC2-DC2	External unique identifiers	Storage; Analytics	Low	
REQ-UC2-DA1	Data science libraries	Storage; Analytics	Medium	
REQ-UC2-DA2	Service SLAs	Storage; Analytics	Medium	

Table 2.2: Use Case 2 Requirements

2.3 Use Case 3: Cloud-based data markets for privacy preserving Consumer analytics (SAP SE)

Use Case 3 represents a business-to-business (B2B) sensitive data sharing scenario in the cloud. Due to the B2B nature Use Case 3 addresses the organizational aspects of data markets. The B2B context differ from plain consumer markets with respect to the fact that not only personal but also sensitive data are of importance. Furthermore, formal security and privacy guarantees for the shared data are of importance to quantify and bound the risk that involved parties accept.

The scenario of this use case particularly addresses two current business objectives: first, the realization of novel data analytics with sanitization; second, achieving formal privacy guarantees for shared data by anonymization. The SAP SE use case includes three actors. Two data owners, a retailer of goods *Retailer* and a producer of goods *Producer* and a cloud platform provider, *SAP SE*, with privacy enhanced services for data sharing. Producer and Retailer desire to increase supply chain insights and redirect marketing budget. Thus Producer and Retailer want to exchange customer experience data (X-Data; e.g., product ratings, historical purchases, etc.) and operational data (O-Data; e.g., Key Performance Indicators such as typical employee salaries per department, bonus payments or sales). SAP SE supports Producer and Retailer with two services: first, local and central functionalities for anonymization of X-Data and O-Data with Differential Privacy; second, tools for interpretation and contextualization of the provided privacy guarantees (e.g., under specific adversaries that strive to re-identify anonymized data). The purpose of this use case is to illustrate how SAP SE cloud platform data owners can use cloud services to enforce sanitization before sharing sensitive or personal data. This provides commercial value to the data owners since privacy risks are lowered and protection of sensitive data are increased.

The mentioned functionalities address the three MOSAICrOWN phases for sanitization as illustrated in Figure 2.7: during data *ingestion* and before data *storage* in the form of local anonymization enforced by Producer or Retailer through services provided by SAP SE before sharing data through the Cloud, and while performing data *analytics* by central anonymization during evaluation of the data analytics query.

Privacy Models. In the *local model* data owners perform data sanitization locally on their own data and send them to a (potentially untrusted) cloud service which aggregates the sanitized data. In the *central model* data owners send their plaintext data to a (trusted) cloud service which aggregates and then sanitizes the data. These two models are presented in Figure 2.8 where the left data

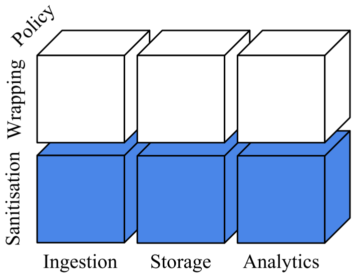

Figure 2.7: Overview of the Use Case 3 dimensions

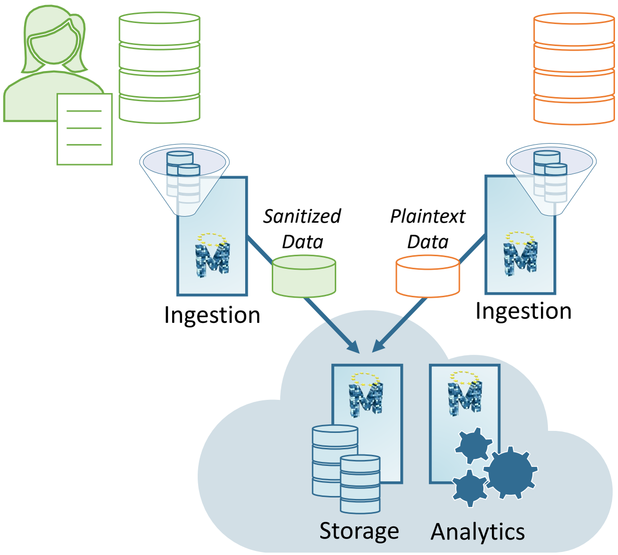

Figure 2.8: Local privacy model, i.e., data owners locally anonymize their data (data owner on the left) and central privacy model, i.e., data owners store plaintext data (data owner on the right) and the cloud platform anonymizes analytical results.

owner locally anonymizes her data (local model) and the right data owner sends her data in plaintext (central model). The local model offers better privacy (the data are not provided in plaintext), but less accuracy. The central model offers better accuracy (the combined data are only sanitized once, e.g., via additive noise), but less privacy (as data are exported to the cloud storage).

Furthermore, a hybrid option exists which simulates the central model in the local model via secure computation, that is, the distributed parties perform the privacy-preserving analytics themselves. Specially crafted secure computation protocols enable data owners to do this without revealing any of their data to each other as visualized in Figure 2.9.

Activity for Local Anonymization. Local anonymization is performed by each data owner already during ingestion. It is useful when a high level of privacy is required, or the data owners have little trust in the operators of the storage and analytics platforms.

The activity follows the left part of the illustration in Figure 2.8: The data owners (e.g., Producer and Retailer) have the original data set that they want to anonymize. Furthermore, they have access to a local anonymization service that will perform the actual sanitization of the data set. The data owners submit the data set to the service in a structured format (such as JSON or CSV), as defined by the service's API. The data owner also might have to specify additional parameters for

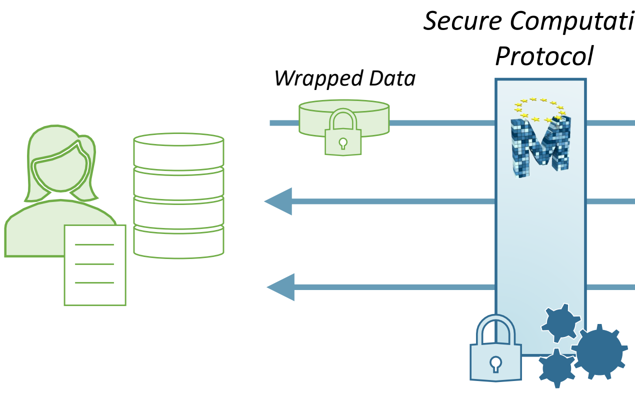

Figure 2.9: Hybrid privacy model.

the anonymization service, such as the actual anonymization mechanism to use and corresponding parameters (e.g. the privacy budget ε that is typically used with differential privacy). The service then anonymizes the submitted data with the desired mechanism and parameters, and returns the anonymization result back to the data owner, again in a format specified by the service API. Finally, the data owner can store the anonymized data set in the cloud, making it available for further analysis or sharing.

Activity for Privacy-Preserving Machine Learning. One activity covering the storage and analytics dimension is privacy-preserving machine learning based on anonymization. Partner SAP SE differentiates privacy-preserving machine learning into two activities. First, a data owner (e.g., Producer) enforces local sanitization during ingestion. Consequently, the data stored by the data owner within the data marketplace are anonymized. If the data owner now desires to learn a machine learning model from the anonymized data, the machine learning will implicitly be anonymized, too. Note that this activity can be extended to multiple data owners (e.g., Producer and Retailer) that collaborate by providing sanitized data to train a machine learning model. Second, a data owner is free to ingest original data. Consequently, the data stored by the data owner within the data marketplace are not sanitized. However, the data owner can attach a policy for sanitization that is evaluated before the data are used for data analytics (e.g., training a machine learning model). If the data owner desires to learn a privacy-preserving machine learning model from the original data, the machine learning training algorithm has to enforce sanitization. Note that this activity again can be extended to multiple data owners that collaborate by providing anonymized data to learn a joined Machine Learning model. The data owners shall be supported in their specific choice of anonymization parameters in both activities (e.g., by assessing risks via re-identification attacks).

Secure storage and data sharing leveraging Trusted Execution Environments. The above described activities provide privacy guarantees based on sanitization of data to data owners. However, security via encryption is required by data owners for data in transit and at rest in addition to privacy to ensure basic information security properties (e.g., confidentiality). Thus an additional activity is required for encryption of data at rest. For this activity the cloud platform provider leverages a Trusted Execution Environment (e.g., Intel SGX processors) at the storage layer in combination with a Public Key Infrastructure that comprises a Certificate Authority (CA). A data owner who wants to store data in the data market contacts the cloud provider CA before ingestion to establish a certificate with a Trusted Execution Environment at the storage layer (e.g., Enclave in Intel SGX). The data owner then encrypts all his data before ingestion. The data owner can furthermore rely on the CA for remote attestation. However, the data owner has to trust the CA.

2.3.1 Functional requirements

Access Control Management. Each data owner is enabled to define and enforce access rights, e.g., via encryption, for his data and stored at the cloud platform. Sharing is enabled by granting access to other parties. Each data owner is supposed to retain control of the encryption key of its data. Keys may be shared for selective access by others.

- REQ-UC3-AC1: Access control decisions should support selective authorizations given by a data owner at different granularity with respect to the authorized subject.
- REQ-UC3-AC2: It should be possible to group several data owners into a group and grant or revoke access for the entire group.
- REQ-UC3-AC3: Each data owner should have its own key generated and kept confidential at its site.
- REQ-UC3-AC4: Key management should support authorizations for groups of data owners.

Sanitization for local functions. Each data owner is enabled to anonymize the data owned by them at time of *ingestion* to the Cloud platform for sharing purposes.

- REQ-UC3-SL1: Anonymization parameters should be chosen by the data owner. The strength of the anonymization parameters is at the discretion of the data owner. This can be achieved by exposing sanitization services through a fine-grained API to data owners.
- REQ-UC3-SL2: Storing the result of the anonymization service (i.e., the sanitized data set) on the cloud should be possible for data owners.
- REQ-UC3-SL3: Sharing the result of the anonymization with data consumers (possibly data owners at the same time) through the cloud platform shall be possible.

Sanitization for central aggregation functions. Each data owner is enabled to anonymize his data while performing *data analytics* on the Cloud platform.

- REQ-UC3-SC1: Anonymization parameters should be chosen by the data owner who executes the central data analytics function. The strength of the sanitization parameters is at the discretion of the data owner. This can be achieved by exposing aggregation functions through a fine-grained API to data owners.
- REQ-UC3-SC2: Collecting inputs from multiple data owners for the aggregation function should be possible.
- REQ-UC3-SC3: Collecting inputs from multiple data owners and evaluation of the aggregation function should be optionally possible via secure computation.
- REQ-UC3-SC4: Anonymization of X-data and O-data should protect the identity of the data subjects and prevent or hinder re-identification attacks.

2.3.2 Non-functional requirements

Performance. The performance of data analysis functionalities under anonymization should remain acceptable from a user perspective.

- REQ-UC3-P1: The anonymization functionalities should be implemented with focus on scalable data processing by the cloud platform owner.
- REQ-UC3-P2: The anonymized data should provide sufficient utility for the desired analytics use cases.

Extendability of sanitization services. The cloud platform shall offer to deploy additional sanitization services for local and central sanitization as state-of-the-art advances.

- REQ-UC3-EX1: The anonymization functionalities should follow an implementation pattern (e.g., share a similar API).
- REQ-UC3-EX2: The cloud platform provider should ensure to provide sound technical means for anonymization from the point of legislation (e.g., following Working Group opinions).

Interpretability for anonymization guarantees. The cloud platform provider should provide a re-identification workbench that provides additional insights on the re-identification protection provided by anonymization services under specific parameters.

- REQ-UC3-IN1: Each anonymization service should be accompanied by a re-identification service that simulates an adversary interested in re-identification of data.

2.3.3 Requirements catalogue

Req. Ref.	**Description**	**Dimension**	**Priority**	**Dependencies on other requirement**
REQ-UC3-AC1	Access control per data owner	Storage	High	REQ-UC3-AC3
REQ-UC3-AC2	Access Control for data owner groups	Storage	Medium	REQ-UC3-AC1, REQ-UC3-AC4
REQ-UC3-AC3	Key per data owner	Storage	High	REQ-UC3-AC1
REQ-UC3-AC4	Keys for data owner groups	Storage	Medium	REQ-UC3-AC3,REQ-UC3-AC2
REQ-UC3-SL1	Local anonymization parameters chosen by the data owner	Ingestion, Sanitization, Policies	High	-
REQ-UC3-SL2	Storing the result of the sanitization service in the Cloud	Ingestion, Sanitization	High	REQ-UC3-SL1, REQ-UC3-AC1

REQ-UC3-SL3	Sharing sanitized results with other data owners	Storage, Policies	Medium	REQ-UC3-SL2, REQ-UC3-AC2
REQ-UC3-SC1	Central anonymization parameters chosen by the data owner	Analytics, Sanitization, Policies	High	-
REQ-UC3-SC2	Collecting inputs from multiple data owners for aggregation	Analytics, Sanitization	Medium	REQ-UC3-SC1
REQ-UC3-SC3	Collecting inputs from multiple data owners via secure computation	Ingestion, Analytics, Sanitization	Medium	REQ-UC3-SC2
REQ-UC3-SC4	Anonymization to protect identity of data subjects and hinder re-identification	Ingestion, Analytics, Sanitization	High	REQ-UC3-SL1, REQ-UC3-SC1
REQ-UC3-P1	Scalability of anonymization functionalities	Ingestion, Analytics, Sanitization	Medium	REQ-UC3-SC2, REQ-UC3-SL2
REQ-UC3-P2	Utility maximizing anonymization functionalities	Ingestion, Analytics, Sanitization	Medium	REQ-UC3-SC2, REQ-UC3-SL2
REQ-UC3-EX1	Extendability pattern for anonymization functions	Ingestion, Analytics, Sanitization	High	REQ-UC3-SL1, REQ-UC3-SC1
REQ-UC3-EX2	Cloud platform provider should provide sound technical means for anonymization	Ingestion, Analytics, Sanitization	High	-
REQ-UC3-IN1	Anonymization services should be accompanied by simulated adversary	Ingestion, Analytics, Sanitization	Medium	-

Table 2.3: Use Case 3 Requirements

3. Requirements Analysis

This chapter describes a synthesis of the common requirements among use cases.

3.1 Synthesis of Common Requirements

Within this section we analyze requirements with respect to the MOSAICrOWN dimensions. There are a number of common topics that run through all MOSAICrOWN use cases. This section seeks to define a common set of terms and concepts to represent the requirements in the use cases, and extract shared elements. All requirements are analyzed along the data life-cycle from ingestion to analytics with respect to the three main dimensions of MOSAICrOWN: Policies, Wrapping, and Sanitization. Furthermore, we extended the analysis by considering Service and Access Control. These categories are explained in the following.

Service. We assume that ingestion, storage and analytics functionalities are eventually realized by individual services (e.g., REST-Service for sanitization of data during ingestion). These service components account for Quality of Service requirements (e.g., latency), but also data management aspects. Thus, we assigned a good portion of requirements that address general requirements to these service categories.

Access Control and Key Management. MOSAICrOWN wants to ensure and provide general information security functionalities not specific to data markets. Thus, we assigned general security functionalities (e.g., PKI) to this category in favor of highlighting only data market specific innovations with respect to Wrapping, Policies and Sanitization.

3.1.1 Ingestion

Table 3.1 groups requirements for the ingestion phase by service, sanitization, wrapping and policies. The table is followed by a description of the categories and their associated common requirement references. The MOSAICrOWN use case prototypes will provide separate service, sanitization, wrapping and policy functionalities for data ingestion. The ingestion requirements are mostly differing between the use cases, yet partly complementary. The granularity with which the ingestion components have been specified by the use case partners varies and we describe differences and common elements in the following.

Category	Use Case Requirements
1) Service	REQ-UC1-DI1, REQ-UC1-DI2, REQ-UC1-DI3, REQ-UC1-DI4, REQ-UC1-AC3, REQ-UC1-DM1,

	REQ-UC1-DM3, REQ-UC1-DM5, REQ-UC1-DE1, REQ-UC1-CQ1, REQ-UC1-CQ2, REQ-UC2-DI1, REQ-UC2-DI2, REQ-UC2-DI3, REQ-UC2-DI4, REQ-UC3-SC3
2) Sanitization	REQ-UC1-P1,REQ-UC1-DI9, REQ-UC2-S1, REQ-UC2-S2, REQ-UC3-SL2, REQ-UC3-EX1, REQ-UC3-EX2, REQ-UC3-IN1, REQ-UC3-SC4, REQ-UC3-SL1, REQ-UC3-P1, REQ-UC3-P2
3) Wrapping	REQ-UC1-DI6, REQ-UC1-P1, REQ-UC2-W1, REQ-UC2-W2, REQ-UC2-W3, REQ-UC2-W4 REQ-UC2-W5, REQ-UC2-W6, REQ-UC2-W7, REQ-UC2-W8, REQ-UC2-W9, REQ-UC2-W10, REQ-UC2-W11, REQ-UC2-W12, REQ-UC2-W13, REQ-UC2-W14, REQ-UC2-W15, REQ-UC2-W16, REQ-UC2-W17
4) Policies	REQ-UC1-DI7, REQ-UC1-DI8, REQ-UC1-DG1, REQ-UC1-DG2, REQ-UC1-DG3, REQ-UC1-DI5, REQ-UC3-SL1

Table 3.1: Categories for use case requirements (ingestion)

Service. UC1 requires data ingestion components close to the data source (e.g., in physical proximity of car sensors). After the corresponding data owner has accepted a license agreement, the data source emits structured and unstructured data either in streams or batches to the service component. Data comprise different data types. User access and key management for encryption in transit to the service, and at rest after processing through the service, and shall be provided by a dedicated key management infrastructure. After ingestions data and access to the data shall be tracked. UC2 requires the service component to derive anonymization needs during ingestion. UC3 requires the service component to specifically support the later application of secure multi-party protocols.

Sanitization. MOSAICrOWN will implement and evaluate a dual approach towards sanitization. For some data, however, sanitization during ingestion will be the default. The use cases address sanitization with varying scopes. UC1 aims for the application of sanitization to mitigate linkage attacks after ingestion. UC2 strives for recognition and isolation of personal data for sanitization. Furthermore, unique outliers should be removed during ingestion. UC3 allows the data owner to enforce anonymization during ingestion. For this purpose one out of a plurality of state-of-the-art anonymization mechanisms shall be selected, and parameterized by the data owner. The data owner should be possibly able to chose to store the sanitized data on the cloud platform or just receive the sanitized data as a service response. UC1 and UC3 require sanitization implementations to run in a scalable manner with low latency.

Wrapping. UC1 requires compression before low latency encryption at rest operations. UC2 requires bucketization of data before applying combinations of sanitization and wrapping, tokenization and fragmentation. UC3 has no formal requirements for wrapping at ingestion.

Policies. Some of the MOSAICrOWN use cases require to manifest exceptions for sanitization or wrapping in the form of policies. In addition, it should be able for data owner's to chose parameters for sanitization and wrapping and ingest these preferences with according data to the service. UC1 permits to define and maintain multiple data governance models in clear language. The governance models should allow the data processor to check data for completeness and preserve identifiers during ingestion. Furthermore, the data governance models should allow a data owner to specify parameters for wrapping and sanitization. UC3 allows data owners to specify and submit anonymization parameters in addition to their data.

3.1.2 Storage

Table 3.2 groups requirements for the storage phase by service, access control & key management, wrapping and policies. Aside from the sanitization and wrapping requirements, the MOSAICrOWN data platform needs to provide functionalities for stored data. Compared to the ingestion functionalities we do observe more synergies as the following analysis of the varying functionalities shows.

Category	Use Case Requirements
1) Service	REQ-UC1-DM1, REQ-UC1-DM2, REQ-UC1-DM3, REQ-UC1-DM4, REQ-UC1-DM5, REQ-UC1-DE1, REQ-UC1-CQ1, REQ-UC1-CQ2
2) Access Control and Key Management	REQ-UC1-AC1, REQ-UC1-AC2, REQ-UC1-AC3, REQ-UC2-AC1, REQ-UC2-AC2, REQ-UC2-AC3, REQ-UC2-AC4, REQ-UC2-AC5, REQ-UC2-AC6, REQ-UC3-AC1, REQ-UC3-AC2, REQ-UC3-AC3
3) Wrapping	REQ-UC2-W1, REQ-UC2-W2, REQ-UC2-W3, REQ-UC2-W4 REQ-UC2-W5, REQ-UC2-W6, REQ-UC2-W7, REQ-UC2-W8, REQ-UC2-W9, REQ-UC2-W10, REQ-UC2-W11, REQ-UC2-W12, REQ-UC2-W13, REQ-UC2-W14, REQ-UC2-W15, REQ-UC2-W16, REQ-UC2-W17

4) Policies	REQ-UC1-DG1, REQ-UC1-DM1, REQ-UC1-DM2, REQ-UC1-DM3, REQ-UC1-DM4, REQ-UC1-DE2, REQ-UC1-DP1, REQ-UC1-DP2, REQ-UC1-DP3, REQ-UC1-DI7, REQ-UC1-DG2, REQ-UC2-DM1, REQ-UC2-DM2, REQ-UC2-DM3, REQ-UC2-DM4, REQ-UC2-DC1, REQ-UC2-DC2

Table 3.2: Categories for use case requirements (storage)

Service. UC1 requires tracking of data access and movement, and enforcement of a license model with regard to data on the platform.

Access Control and Key Management. All MOSAICrOWN use cases should provide the specification of access controls and authorizations on varying granularity. Furthermore, all use cases require a centralized key management infrastructure. UC2 limits data access to APIs, and data exports to aggregates or statistical models. UC3 allows data owners to specify access controls.

Wrapping. UC2 requires wrapping, tokenization and fragmentation of data at the storage level.

Policies. UC1 requires data governance support also at the storage level. Data shall be updatable, but identifiers shall still be preserved. It should be possible to distinguish between data for data sharing and data analytics. Guarantees for deletion of data shall be obtainable. UC2 requires controller-to-controller contracts to obtain data controller rights when data are stored in the platform, and to allow combination of data. Furthermore, repository external unique identifiers shall be substituted with repository specific unique identifiers.

3.1.3 Analytics

Table 3.3 groups requirements for the analytics phase by service, sanitization and policies. Regarding analytics, UC1 is mainly requiring data governance functionalities, UC2 is specifying platform requirements, and UC3 is demanding privacy-preserving analytics functions that enforce anonymization during computation.

Category	**Use Case Requirements**
1) Service	REQ-UC1-DM5, REQ-UC1-DP1, REQ-UC1-DP3, REQ-UC1-DE1, REQ-UC1-CQ1, REQ-UC1-CQ2, REQ-UC1-AC3, REQ-UC1-DM1, REQ-UC2-DA1, REQ-UC2-DA2

2) Sanitization	REQ-UC1-DP2, REQ-UC1-P1, REQ-UC1-P2, REQ-UC2-S5, REQ-UC3-SC2, REQ-UC3-SC3, REQ-UC3-SC4, REQ-UC3-EX1 REQ-UC3-EX2, REQ-UC3-IN1, REQ-UC3-SC1, REQ-UC3-P1, REQ-UC3-P2
3) Policies	REQ-UC1-DE2, REQ-UC1-DG1, REQ-UC1-DG2, REQ-UC1-P2, REQ-UC2-S2, REQ-UC2-JD1, REQ-UC2-JD2, REQ-UC2-JD3, REQ-UC3-SC1

Table 3.3: Categories for use case requirements (analytics)

Service. UC1 requires the data platform to also track access to and movement of data on the analytics level. Furthermore, the analytics functionalities have to have access to the key management infrastructure for reading, and writing data. UC2 requires the data platform to support standard data science libraries and to support SLAs.

Sanitization. Both UC1 and UC3 require that analytics output must be sanitized and provide meaningful utility. UC3 additionally requires support for analytics via cryptographic tools, i.e., secure computation. Furthermore, the data analytics service should provide functionalities for anonymization during analytics computation and shall be accompanied by a re-identification service to assess anonymization parameters.

Policies. UC1 requires a clear separation between data providers and data analytics. Both UC1 and UC2 demand that sanitization requirements for analytics shall be specified upfront. UC1 desires a data governance model that was specified beforehand. UC2 desires that the data owner can specify anonymization parameters.

4. Use Case Requirements Comparison to Current Technologies

This chapter provides a comparison between the use case requirements and state-of-the-art for security and privacy in data markets. The analysis is motivated by the wish to identify concrete gaps in current technologies in areas that MOSAICrOWN investigates. MOSAICrOWN will then advance state-of-the-art with technical solutions for data protection. The solutions will be deployed over the course of the project in the context of the different use cases.

4.1 Use Case 1

The primary challenge in Use Case 1 is the management of fine-grained, dynamic access controls and roles in a shared data market. Solutions for encryption and anonymization already exist that are sufficient for 1:1 relationships between data owners and data consumers, including lightweight solutions that are well suited to applications with resource constraints or latency sensitivity [SSMP17] and those which maintain the ability to search the underlying data without decryption [BHJP14]. Similarly, key management protocols to support encryption and access control policies are well documented, e.g. [DHFS19]. However, these existing solutions do not appropriately or efficiently deal with multi-user scenarios on a shared platform. Nor do they account for the involvement of third parties, whether they be the cloud provider hosting the platform, or the security provider managing keys, encryption processes, or anonymization processes [AAUC18]. The crucial aspect of this use case is that data be shared and analyzed openly, but without risking privacy breaches by allowing access to third parties or legitimate consumers on the data market who lack the access rights to specific data and also without risking the safety of the users from malicious use of the data.

Therefore, the solutions that will result from MOSAICrOWN will support the protection of data close to source to minimize the risk to sensitive data. They will also advance the state of the art in anonymization techniques to retain as much as possible the utility of the data. Novel encryption schemes, such as selective encryption, traditionally used for image data [GC15], will also push the state of the art, offering an efficient approach to selectively sharing and analyzing data with different data consumers. Consideration to lightweight protocols suitable to automotive security [JWLZ18] should also be taken into account.

Regarding the real life automotive threats [Mil19], special consideration should be given to any possible automotive exploits when facilitating access control to automotive data. Such attacks depend on large amounts of automotive diagnostic data and data sanitization will reduce these risks.

4.2 Use Case 2

One gap addressed by Use Case 2 is the lack of ready to use wrapping and sanitization standards for data combination scenarios where there are multiple data owners and data users. Data owners may currently use a variety of data wrapping techniques that are recommended through NIST: from encryption to protect in-transit and at-rest data, to pseudo-anonymization techniques for adding noise to their stored data, to hashing for additional layers of security, there are multiple techniques suitable for data protection for the various needs of different organizations [NIS16]. However, while the NIST recommendations will provide adequate protection to all forms of data, there is constant research being conducted on the future of data wrapping techniques. There is increasing interest in blockchain technologies to protect data, especially as Cloud and Internet-based storages systems gain traction [TDD$^+$19]; however, it is not yet state-of-the-art and still being researched. By continually researching and building a greater understanding around the potential of blockchain and publicly stored yet protected data (like the Cloud), the data security community has placed an increased focus on preparing and predicting future needs before a high-stakes data breach occurs.

Based on the current state of the field and NIST recommendations, the recommendations made in Use Case 2 are on par, if not stronger, than what is recommended. While Use Case 2 does not suggest any forms of blockchain technology, as it has not yet been fully recommended by NIST, it does suggest various implementation tactics dependent upon use case. For example, certain encryption algorithms, like the RSA, are suggested only when exchanging keys over the internet [VD19]. Understanding the nuances of these recommendations and their use cases requires further time and investigation. This discussion and further details will be discussed in WP4.

A major differentiator of Use Case 2 is a state-of-the-art data wrapping layer that will enable the data contributor or MOSAICrOWN the option to select which combination of wrapping techniques they would like applied to their data, down to a column level. This process starts with determining the relative sensitivity of the data elements provided and identifying the types of wrapping required to ensure enough data protection. Based upon this review, the system will provide the flexibility to select multiple data wrapping techniques to be applied to the data, instead of just a standard encryption or anonymization wrapping. The flexibility and variety offered from the system, through a catalogue, will provide a secondary level of security from an outward attack. The system will be built to handle any combination of wrapping techniques such as pseudo-anonymization, encryption, hashing and tokenization. All of these techniques, in use together, significantly reduce the risks associated with data processing, while maintaining the data's utility [Sec10].

4.3 Use Case 3

Several approaches have been proposed for regulating access and protecting data possibly stored and managed by external parties outside the data owner control (e.g., [BDF$^+$16]). Most of these solutions are based on the adoption of client-side encryption for protecting data confidentiality. For example, SAP SE designed SEEED (SEarch-OvEr-Encrypted-Data) [G$^+$14, HKD15], a database system that utilizes client-side encryption techniques to support query processing over encrypted data.

In contrast, Use Case 3 focuses on sanitization techniques that potentially complement encryption based functionalities. A major differentiator of Use Case 3 is the focus on sanitization via noise addition (e.g., differential privacy [Dwo06]). The loss of utility in a differentially private data set depends on the privacy parameter ε and the sensitivity of the anonymized functions, and some models have been proposed to provide a rational method for choosing ε (e.g., [CT13, HGH$^+$14, LC11]). However, no adoption of such models took place yet for large scale deployments such as those found in an enterprise context.

Enforcement of differential privacy at the data source, that is, local differential privacy, has been addressed before in the context of releasing geo-coordinates for location based services (e.g., [ABCP13]) and usage statistics about web browsers (e.g., [EPK14]). Another decentralized approach called *federated learning* [MMR$^+$17] can also benefit from differential privacy to achieve stronger privacy guarantees [GKN17, MRTZ18]. Use Case 3 will design novel local differential privacy mechanisms to support X-Data and O-Data, considering individual data sources, with the goal of achieving scalability and sufficient utility for the designated use cases.

Recently, to improve utility, differential privacy has been combined with secure computation. Within these lines of work secure computation is commonly realized via homomorphic encryption or secret sharing, but also more general garbled circuits, e.g., [GKM$^+$16, GX17, HMFS17, RN10]. Moreover, secure computation can also be used to strengthen the privacy guarantees as e.g. in federated learning [BIK$^+$16].

5. Conclusions

This deliverable presented each of the three use cases of MOSAICrOWN identified by the industry partners, along with actors per use case and their activities. The use cases span from privacy in data markets for IoT data over financial data markets to enterprise data markets.

By performing a detailed analysis of functional and non-functional requirements per use case we could identify different technological requirements. Although the three MOSAICrOWN use cases address different aspects of data markets we observed that a non-negligible portion of requirements is shared between all use cases per enforcement technique by also performing a synthesis of requirements.

Furthermore, we made sure that each use case advances state-of-the-art by identifying the gaps we see between data market requirements and key research results throughout the past. From the requirements and state-of-the-art analysis we conclude the following.

- *The use cases cover multiple deployment scenarios to fit a wide variety of applications.* No one-size-fits-all solution for different industry settings or deployment scenarios exists, thus, MOSAICrOWN supports multiple deployment options by considering where the protection techniques should be applied: Data sanitization and wrapping *locally* at the data source, a *central* platform providing these functionalities as a service during ingestion and the entire data life-cycle, as well as *hybrid* scenarios, combining both previously mentioned scenarios, which can be augmented by cryptographic tools (e.g., secure computation).

- *Wrapping and sanitization provide protection beyond policy-based controls.* To meet security and privacy requirements for the use cases MOSAICrOWN provides different wrapping as well as data sanitization techniques to augment policy-based protection mechanisms which control data access, usage and sharing. Wrapping is typically reversible if provided a key (e.g., encryption), whereas data sanitization is a non-reversible transformation (e.g., differential privacy).

- *Analytics for wrapped or sanitized data provides additional technical challenges.* To provide meaningful utility while providing strong security and privacy guarantees special care has to be taken with regards to how to sanitized the data to protect the privacy of individuals while allowing statistical inference. For wrapping techniques key management and enforcement of policies are vital to provide protection and yet enable authorized participants to perform analytics over the protected data.

Bibliography

[AAUC18] Abbas Acar, Hidayet Aksu, A. Selcuk Uluagac, and Mauro Conti. A survey on homomorphic encryption schemes: Theory and implementation. *ACM Comput. Surv.*, 51(4):79:1–79:35, July 2018.

[ABCP13] Miguel E. Andrés, Nicolás E. Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. Geo-indistinguishability: differential privacy for location-based systems. In *Proc. of CCS 2013*, Berlin, Germany, November 2013.

[BDF$^+$16] Enrico Bacis, Sabrina De Capitani di Vimercati, Sara Foresti, Stefano Paraboschi, Marco Rosa, and Pierangela Samarati. Mix&slice: Efficient access revocation in the cloud. In *Proc. of CCS 2016*, Vienna, Austria, October 2016.

[BHJP14] Christoph Bösch, Pieter Hartel, Willem Jonker, and Andreas Peter. A survey of provably secure searchable encryption. *ACM Comput. Surv.*, 47(2):18:1–18:51, August 2014.

[BIK$^+$16] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for federated learning on user-held data. *CoRR*, abs/1611.04482, 2016.

[CT13] Chris Clifton and Tamir Tassa. On syntactic anonymity and differential privacy. *TDP*, 6(2):161–183, 2013.

[DHFS19] Laurin Doerr, Michael Heigl, Dalibor Fiala, and Martin Schramm. Comparison of energy-efficient key management protocols for wireless sensor networks. In *Proc. of IECC 2019*, Okinawa, Japan, July 2019.

[Dwo06] Cynthia Dwork. Differential privacy. In *Proc. of ICALP 2006*, Venice, Italy, July 2006.

[EPK14] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proc. of CCS 2014*, Scottsdale, AZ, USA, November 2014.

[G$^+$14] Patrick Grofig et al. Experiences and observations on the industrial implementation of a system to search over outsourced encrypted data. In *Proc. of Sicherheit*, Vienna, Austria, March 2014.

[GC15] Danilo De Oliveira Gonçalves and Daniel G. Costa. A survey of image security in wireless sensor networks. *J. Imaging*, 1(1):4–30, June 2015.

[GKM$^+$16] Vipul Goyal, Dakshita Khurana, Ilya Mironov, Omkant Pandey, and Amit Sahai. Do distributed differentially-private protocols require oblivious transfer? In *Proc. of ICALP 2016*, Rome, Italy, July 2016.

[GKN17] Robin C. Geyer, Tassilo Klein, and Moin Nabi. Differentially private federated learning: A client level perspective. *CoRR*, abs/1712.07557, 2017.

[GX17] Slawomir Goryczka and Li Xiong. A comprehensive comparison of multiparty secure additions with differential privacy. *IEEE TDSC*, 14(5):463–477, October 2017.

[HGH$^+$14] JustinKhanna Hsu, Marco Gaboardi, Andreas Haeberlen, Sanjeev Khanna, Arjun Narayan, and Benjamin C Pierce. Differential privacy: An economic method for choosing epsilon. In *Proc. of CSF 2014*, Vienna, Austria, July 2014.

[HKD15] Isabelle Hang, Florian Kerschbaum, and Ernesto Damiani. ENKI: Access control for encrypted query processing. In *Proc. of SIGMOD 2015*, Melbourne, Victoria, Australia, May–June 2015.

[HMFS17] Xi He, Ashwin Machanavajjhala, Cheryl Flynn, and Divesh Srivastava. Composing differential privacy and secure computation: A case study on scaling private record linkage. In *Proc. of CCS 2017*, Dallas, TX, USA, October–November 2017.

[JWLZ18] Ahmer Khan Jadoon, Licheng Wang, Tong Li, and Muhammad Azam Zia. Lightweight cryptographic techniques for automotive cybersecurity. *Wireless Communications and Mobile Computing*, 2018, 2018.

[LC11] Jaewoo Lee and Chris Clifton. How much is enough? choosing ε for differential privacy. In *Proc. of ISC 2011*, Xi'an, China, October 2011.

[Mil19] Charlie Miller. Lessons learned from hacking a car. *IEEE Design Test*, 36(6):7–9, December 2019.

[MMR$^+$17] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proc. of AISTATS*, Fort Lauderdale, FL, USA, April 2017.

[MRTZ18] H. Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning differentially private recurrent language models. In *Proc. of ICLR 2018*, Vancouver, BC, Canada, April-May 2018.

[NIS16] NIST. Standards incorporated by reference. https://www.nist.gov/standardsgov/what-we-do/federal-policy-standards/sibr-database, 2016.

[RN10] Vibhor Rastogi and Suman Nath. Differentially private aggregation of distributed time-series with transformation and encryption. In *Proc. of SIGMOD 2010*, Houston, TX, USA, June–July 2010.

[Sec10] Townsend Security. Tokenization: A cost-effective and easy path to compliance and data protection. *Whitepaper by Townsend*, 2010.

[SSMP17] Saurabh Singh, Pradip Kumar Sharma, Seo Yeon Moon, and Jong Hyuk Park. Advanced lightweight encryption algorithms for iot devices: survey, challenges and solutions. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–18, May 2017.

[TDD$^+$19] Paul J Taylor, Tooska Dargahi, Ali Dehghantanha, Reza M Parizi, and Kim-Kwang Raymond Choo. A systematic literature review of blockchain cyber security. *Digital Communications and Networks*, 2019.

[VD19] Gahan A V and Geetha D. Devanagavi. A empirical study of security issues in encryption techniques. *International Journal of Applied Engineering Research*, 14(5):1049–1061, 2019.